



Training Data Lab

Aplicaciones de data mining, modelamiento y machine learning
para ciencias sociales

Bastián González-Bustamante

University of Oxford

Universidad de Santiago de Chile

✉ bastian.gonzalezbustamante@politics.ox.ac.uk

Presentación preparada para el primer taller del laboratorio de datos

Training Data Lab, 20 de abril de 2021

Hoja de ruta

1. Introducción
2. Proyectos realizados
3. Proyectos en curso
4. Próximos proyectos
5. Tufte Working Papers
6. Lineamientos de colaboración



Introducción

Introducción

© 2020 **Training Data Lab** es un grupo de investigación que se enfoca en aplicaciones de ciencia de datos en ciencias sociales en tres áreas interconectadas: **minería de datos, modelamiento econométrico y aprendizaje automático**. Por una parte, buscamos recoger datos con técnicas de minería para elaborar modelos econométricos con técnicas observacionales o de emparejamiento.

Por otro lado, nos enfocamos en entrenar modelos con técnicas de aprendizaje automático y profundo etiquetando conjuntos de datos para diferentes proyectos. Lo anterior, nos permite clasificar datos no codificados usando nuestros modelos entrenados incorporando validación humana en el flujo de trabajo, lo que mejora la inteligencia artificial en los procesos de aprendizaje.

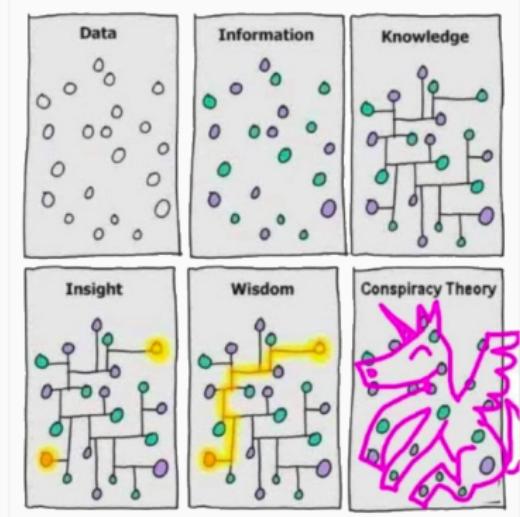
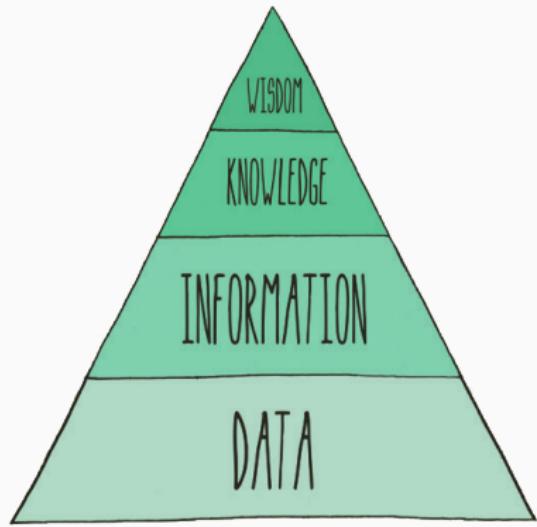


Universiteit
Leiden



UNIVERSIDAD
MAYOR

Introducción



Proyectos realizados

Rastreador Online COVID-19

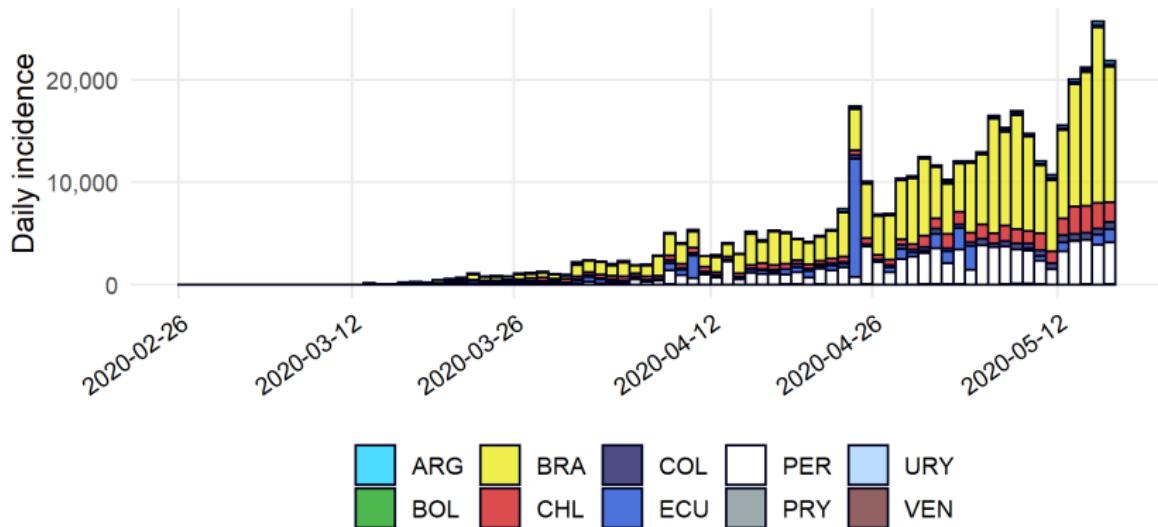
-  Bastián González-Bustamante (responsable)
-  bgonzalezbustamante.github.io/COVID-19-South-America

COVID-19 in South America Tracker. Rastreador online para Sudamérica desplegado entre el 11 de marzo y mediados de mayo de 2020 con datos de JHU. Este tracker presentaba periódicamente las curvas epidémicas en función de la incidencia en los distintos países de la región, el R estimado, la distribución de intervalo de serie explorada (SI) y simulaciones de incidencia futura. Además, se ofrecían comparaciones con algunos países europeos.

Early Government Responses to COVID-19 in South America. Artículo publicado en World Development (González-Bustamante, 2021), código en R disponible en el repositorio y archivo completo de replicación en Elsevier que integra datos de JHU, Oxford, PAHO, V-Dem y WB.

Rastreador Online COVID-19

COVID-19 - Coronavirus Epidemic Curve in South America

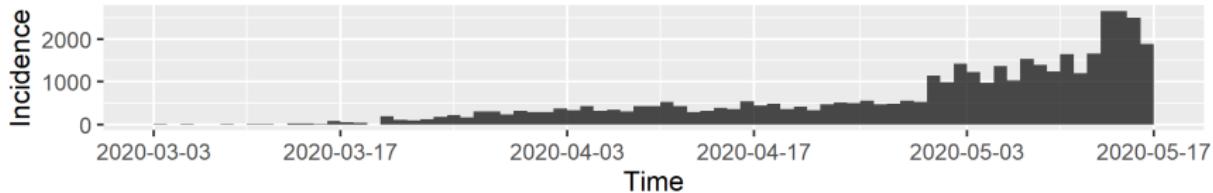


Data up to 16 May - DOI: 10.17605/OSF.IO/Y6C7Z

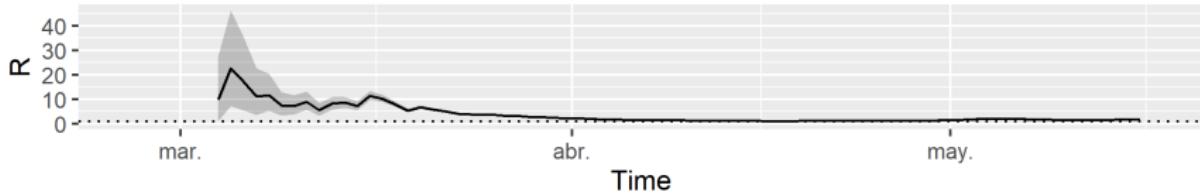
Note: There is inconsistency on 12 April data in the Uruguayan case.
As well as on 7, 8, and 9 May data in the Ecuadorian case.

Rastreador Online COVID-19

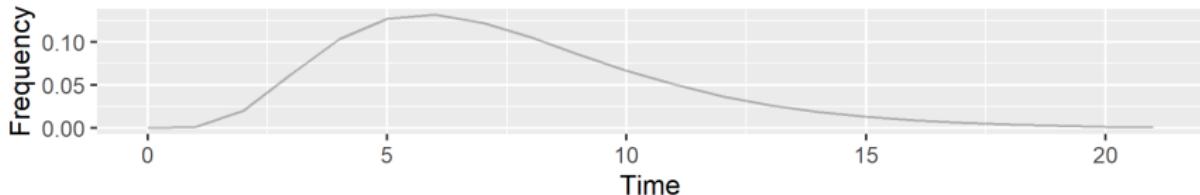
Epidemic curve



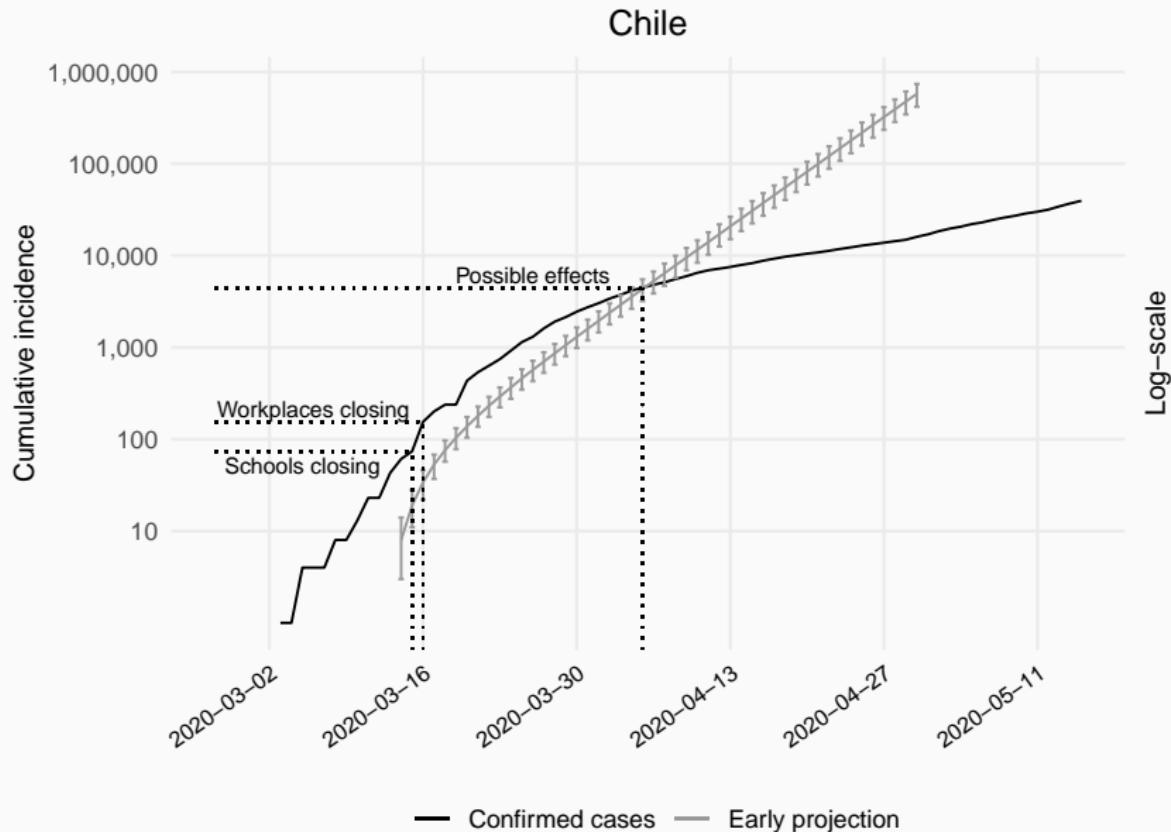
Estimated R



Explored SI distribution



Rastreador Online COVID-19



Algoritmo OCR para servicio civil chileno

 Bastián González-Bustamante, Matías Astete y Berenice Orvenes
(responsables)

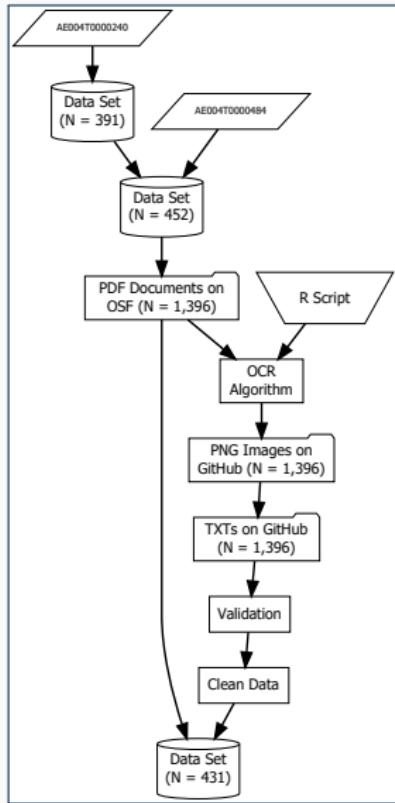
 DOI: [10.17605/OSF.IO/WBF6M](https://doi.org/10.17605/OSF.IO/WBF6M)

 training-datalab.com/projects/chilean-civil-service

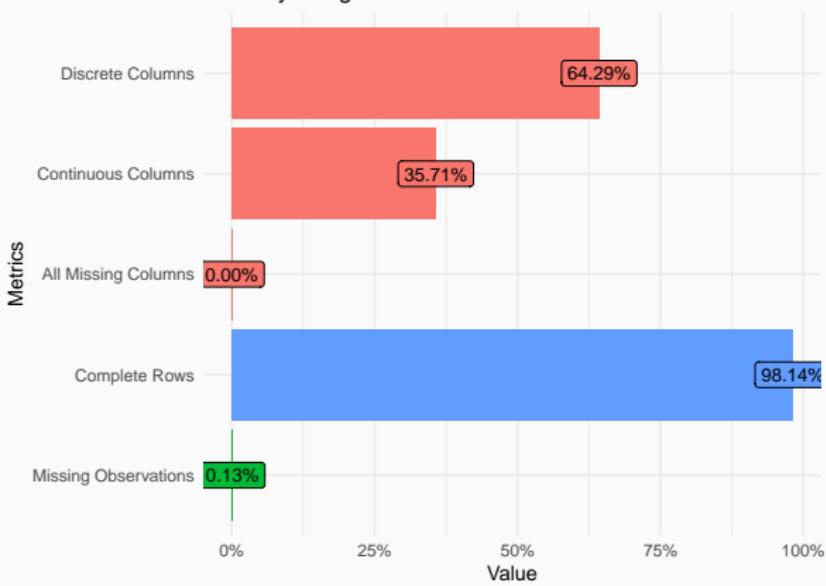
A Novel Dataset on Members of the Chilean Civil Service. Este conjunto de datos contiene información detallada de 431 altos directivos públicos del primer nivel jerárquico del servicio civil chileno durante el período 2009-2017. Fue creado con dos solicitudes de acceso a información pública realizadas a la DNSC y una revisión de 1.396 documentos públicos, principalmente decretos y noticias institucionales. Estos documentos fueron digitalizados con algoritmos de minería de datos y revisados de forma semi-automatizada exhaustivamente.

 Revisar el [preprint en SocArXiv](#) y el [artículo publicado en español](#).

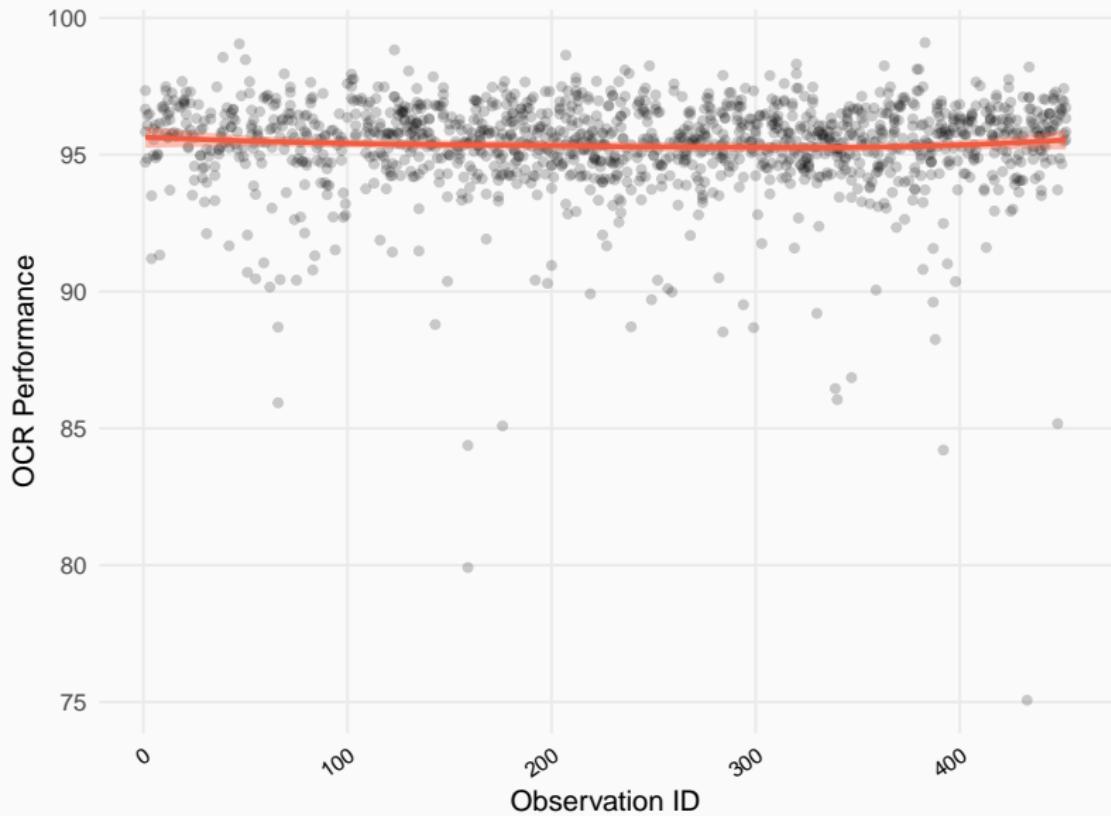
Algoritmo OCR para servicio civil chileno



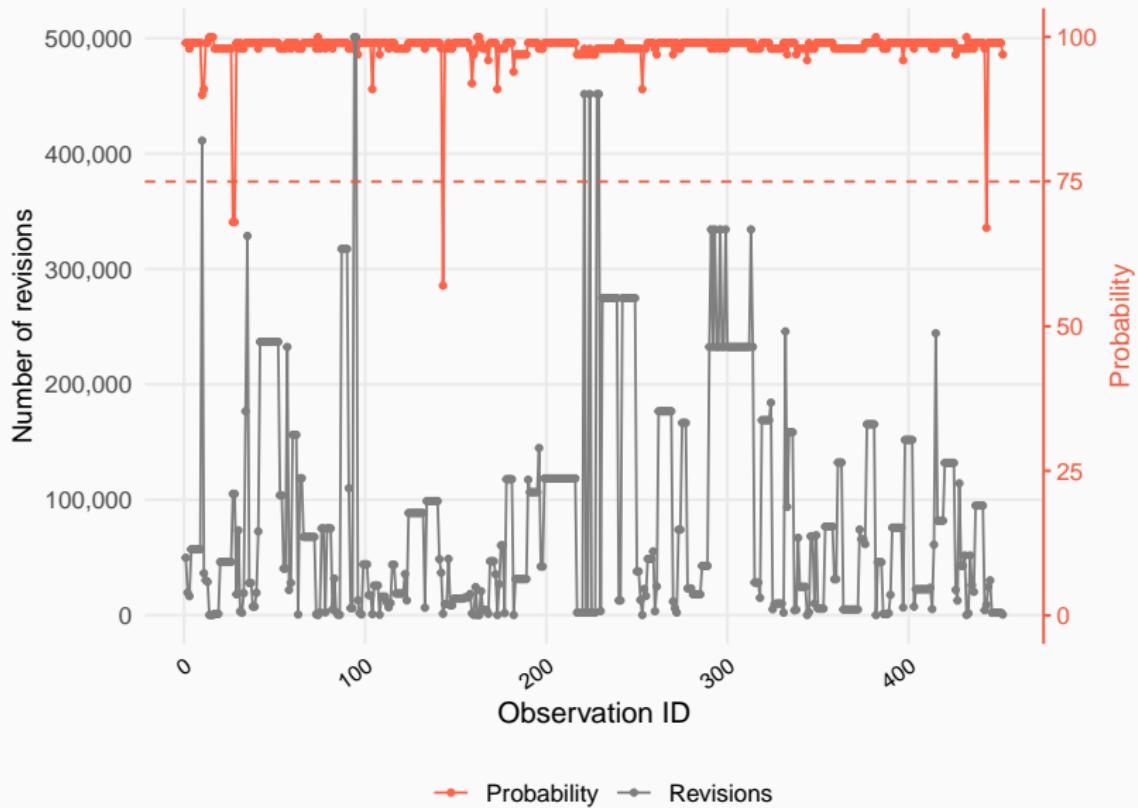
Memory Usage: 243 Kb



Algoritmo OCR para servicio civil chileno



Algoritmo OCR para servicio civil chileno



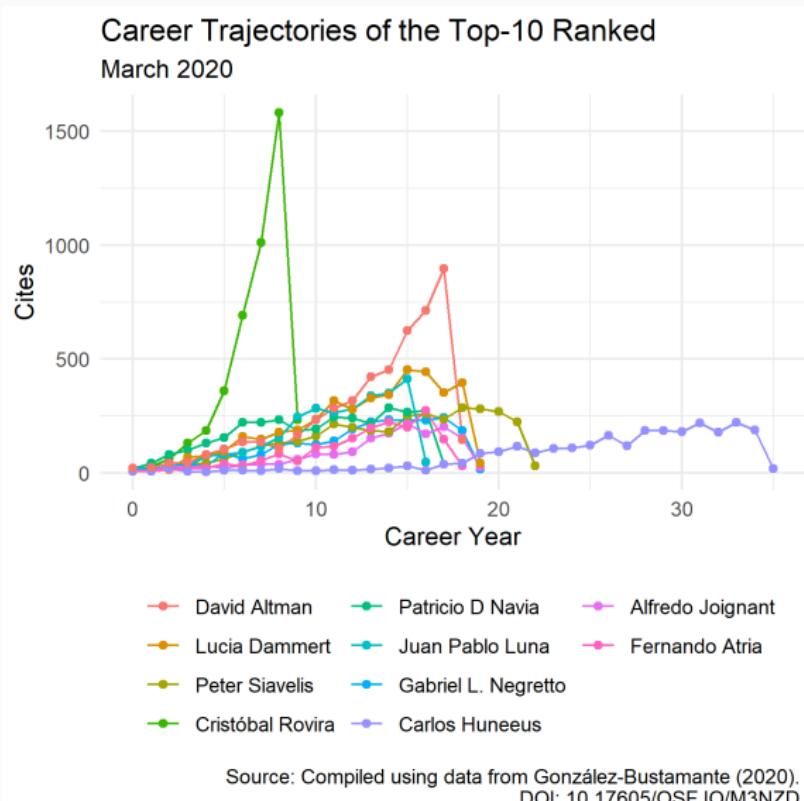
Proyectos en curso

Minería de datos en Google Scholar

-  Bastián González-Bustamante (responsable)
-  Alejandro Olivares, Carla Cisternas y Rodrigo Cuevas (colaboradores)
-  bgonzalezbustamante.com/cps-ranking

Chilean Political Science Ranking (CPS-Ranking). El ranking presenta una medición trimestral del H-Index y total de citas de diversos investigadores. Fue elaborado con todos quienes participaron en el congreso de la ACCP en 2018 que poseían una cuenta activa en  Google Scholar. Posteriormente, se han incorporado investigadores por sugerencia de otros que son parte del ranking. En la primera medición (diciembre, 2019) fueron removidos quienes presentaban problemas de autoría. Luego no se han removido investigadores, pero en la medición de marzo de 2021 ofrecemos un índice de consistencia.

Minería de datos en Google Scholar

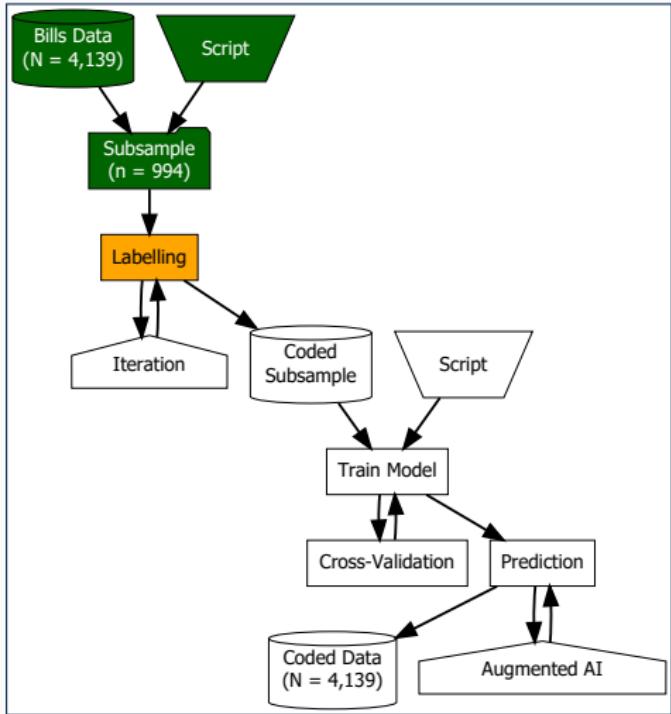


Algoritmo clasificador para mociones legislativas

-  Carla Cisternas y Bastián González-Bustamante (responsables)
-  Diego Aguilar (colaborador)
-  **Estamos reclutando colaboradores y ayudantes**
-  training-datalab.com/projects/chilean-congress-bills

Training Data on Chilean Congress Bills. A partir de un conjunto de datos de proyectos de ley de la Cámara de Diputados de Chile entre 2006 y 2018 ($N = 4.139$), período que corresponde a tres administraciones, extraemos una submuestra aleatoria considerando algunos proyectos de ley por mes. En esta submuestra realizamos dos procedimientos de codificación de datos para identificar tanto el tema del proyecto de ley como su alcance territorial.

Algoritmo clasificador para mociones legislativas



Deep learning para clasificar publicaciones

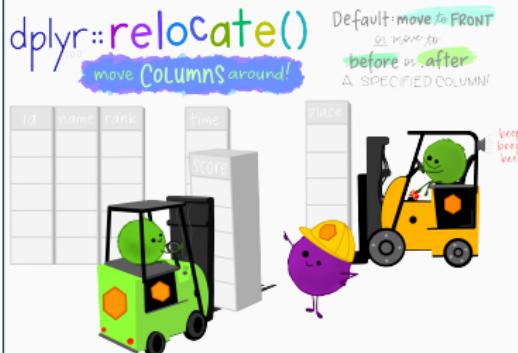
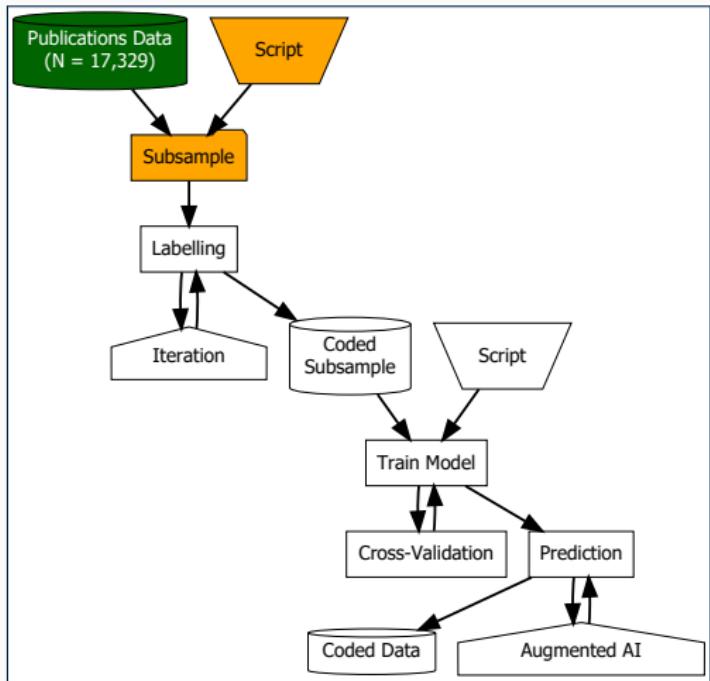
 Bastián González-Bustamante, Alejandro Olivares y Carla Cisternas
(responsables)

 training-datalab.com/projects/political-science-publications

Algoritmos de aprendizaje profundo para clasificar la producción científica: Evidencia de la ciencia política en español. Este trabajo presenta los resultados de la primera aplicación de un proceso de entrenamiento de datos y algoritmos para clasificar la productividad de la ciencia política publicada en español en revistas indexadas en el Social Sciences Citation Index de Web of Science (SSCI-WoS) y Scopus.

 Revisar el abstract en este [repositorio](#) (no distribuir por favor).

Deep learning para clasificar publicaciones



Próximos proyectos

Proyecto VIP de machine learning

- 👤 Carla Cisternas y Francisco Castañeda (responsables)
- 👤 Bastián González-Bustamante, Rodrigo Cuevas y Diego Aguilar (colaboradores)
- 🏠 training-datalab.com/projects/vip-project-machine-learning

Aplicaciones de Machine Learning en políticas públicas y economía.

Este proyecto VIP busca conformar un equipo multidisciplinario de investigación verticalmente integrado para realizar diferentes aplicaciones de aprendizaje automático en temas de políticas públicas y economía. En esta primera versión, nos centramos en las dinámicas de desinformación en temas económicos y de interés público.

Posteriormente, en futuras versiones, esperamos abordar y analizar otros fenómenos relacionados con la formulación de políticas públicas sectoriales y regulación de mercados.

Proyecto VIP de machine learning



APLICACIONES DE MACHINE LEARNING EN POLÍTICAS PÚBLICAS Y ECONOMÍA

Scraper twConstitution

-  Bastián González-Bustamante (responsable)
-  Estamos reclutando colaboradores y ayudantes
-  bgonzalezbustamante.github.io/twConstitution
-  Pronto disponible en training-datalab.com

Twitter Online Tracker of the Chilean Referendum for a New Constitution. Rastreador online durante el plebiscito para una nueva Constitución en octubre de 2020. Contiene datos diarios de #Apruebo y #Rechazo entre el 26 de septiembre y 01 de noviembre ($N = 2.529.134$). Algunas variables disponibles son fecha, hora, usuario, texto, recuento de RTs y favs, ubicación, etc. Es necesario fusionar los conjuntos diarios y limpiar los datos.

El objetivo es entrenar un algoritmo clasificador y realizar un benchmarking de mediciones de emotividad.

Scraper twConstitution

twConstitution



Twitter Online Tracker of the Chilean Referendum for a New Constitution

[View the Project on GitHub](#)
bgonzalezbustamante/twConstitution

Twitter Online Tracker of the Chilean Referendum for a New Constitution

version v1.2.6 issues 1 open issues 4 closed DOI 10.17605/OSF.IO/73NDB

license CC-BY-4.0 made with R v4.0.2 made with Jekyll

This is a Twitter online tracker of the Chilean referendum for a new Constitution in October 2020, which contains daily datasets on **#Apruebo** (see words network, *forthcoming*) and **#Rechazo** (see words network, *forthcoming*) viewpoints on this social media.

Data sets are scraped and uploaded regularly. Some of the variables are date, hour, username, tweet text, RT count, fav count, location, among others. The data was collected during the afternoon each day. A couple of exceptions, such as October 6th and 26th, were collected early morning on the following day. This is not an issue because the data could be sliced, and in order to work with the whole period, it is necessary to merge the sets and retain unique cases.

Dataset	Date	Year	N	Size	Format
#Apruebo	Nov. 01	2020	45,195	46.5 MB	CSV
#Rechazo	Nov. 01	2020	22,142	24.0 MB	CSV
#Apruebo	Oct. 31	2020	50,244	52.1 MB	CSV
#Rechazo	Oct. 31	2020	27,331	29.9 MB	CSV
#Apruebo	Oct. 30	2020	53,889	56.1 MB	CSV
#Rechazo	Oct. 30	2020	33,479	36.6 MB	CSV

Tufte Working Papers

Tufte Working Papers

-  Bastián González-Bustamante (editor)
-  Elinor Luco (asistente editorial)
-  training-datalab.com/tufte-working-papers

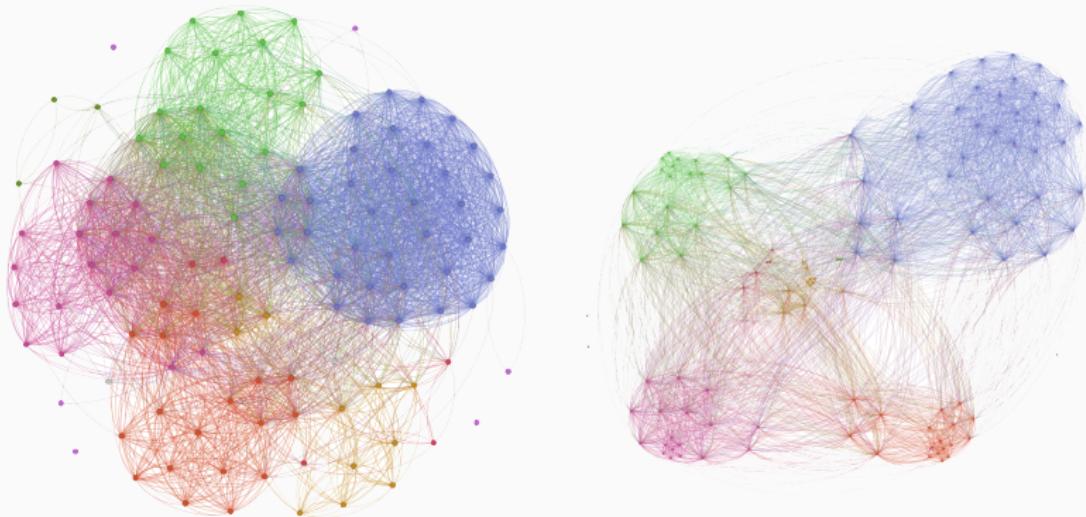
© 2020 **Tufte Working Papers (ISSN 2735-6043)** es una publicación continua basada en Training Data Lab con el apoyo logístico de la Universidad de Santiago de Chile. Esta serie de documentos promueve el debate en ciencias sociales, especialmente en temas relacionados con la ciencia política y las políticas públicas. La serie incluye trabajos inéditos y versiones revisadas de publicaciones previas que proponen **técnicas de investigación innovadoras u ofrecen información empírica novedosa**. Los trabajos se publican en español e inglés, aunque se privilegian las publicaciones en español con el fin de apoyar la difusión del conocimiento de acceso abierto en Iberoamérica.

Esta serie es editada en **Tufte-LaTeX**, una plantilla LaTeX inspirada por Edward R. Tufte. LaTeX es un software libre que permite la composición de textos con alta calidad tipográfica. Por otro lado, cuenta con una **política ética y normas de estilo** basadas en diversas declaraciones e iniciativas (Singapur, COPE, Budapest, Bethesda y Berlín). En consecuencia, los trabajos se licencian bajo Creative Commons.

La serie utiliza la **taxonomía CRediT** para identificar las contribuciones exactas de cada investigador en caso de coautorías o asistencias de investigación.

Se utiliza una **revisión abierta** con uno o dos expertos en la que se divultan las identidades de los autores y árbitros.

Aplicación de ForceAtlas2, un algoritmo de diseño gráfico continuo, para el estudio de las élites. Working paper enfocado en análisis de redes sociales y su visualización ([González-Bustamante y Cisternas, 2020](#)).



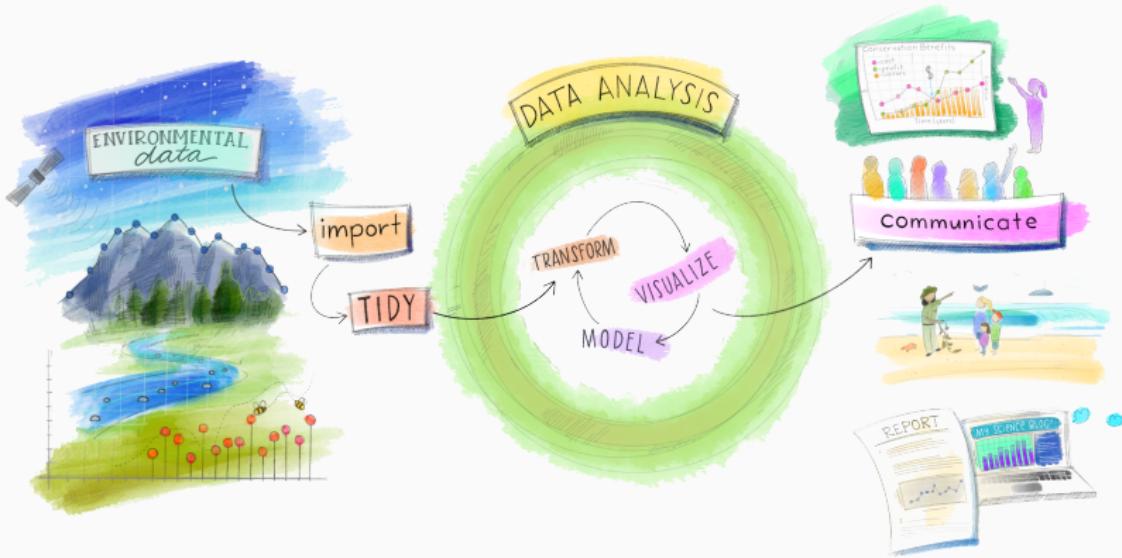
Lineamientos de colaboración

Lineamientos de colaboración

Taxonomía CRediT. Es una taxonomía de alto nivel, que incluye 14 roles, que se puede utilizar para representar los roles que suelen desempeñar los contribuyentes a la producción científica académica. Los roles describen la contribución específica de cada colaborador a la producción académica. Más detalles en casrai.org/credit y bgonzalezbustamante.com/credit.



Lineamientos de colaboración



Solo un poco de información adicional...

Contacto

Bastián González-Bustamante

DPhil (PhD) Researcher

Department of Politics and International Relations

& St Hilda's College

University of Oxford

📍 St Hilda's College, Cowley Place, Oxford OX4 1DY

✉ bastian.gonzalezbustamante@politics.ox.ac.uk

🏠 <https://bgonzalezbustamante.com>

Profesor Instructor

Departamento de Gestión y Políticas Públicas

Facultad de Administración y Economía

Universidad de Santiago de Chile

📍 Av. Lib. B. O'Higgins 3363, Estación Central, Santiago

✉ bastian.gonzalez.b@usach.cl



Presentación compilada con **LATEX** y algunos

🔗 Descargar la versión más reciente desde [GitHub](#)

🎥 Descargar el video de la presentación desde [Dropbox](#)

🕒 Artwork utilizado disponible en [GitHub](#)

Muchas gracias por su atención

